

1 Fitness tracking reveals task-specific associations
2 between memory, mental health, and exercise

3 Jeremy R. Manning^{1, *}, Gina M. Notaro^{1,2}, Esme Chen¹, and Paxton C. Fitzpatrick¹

4 ¹Dartmouth College, Hanover, NH

5 ²Lockheed Martin, Bethesda, MD

6 *Address correspondence to jeremy.r.manning@dartmouth.edu

7 October 11, 2021

8 **Abstract**

9 Physical exercise can benefit both physical and mental well-being. Different forms of exercise
10 (i.e., aerobic versus anaerobic; running versus walking versus swimming versus yoga; high-
11 intensity interval training versus endurance workouts; etc.) impact physical fitness in different
12 ways. For example, running may substantially impact leg and heart strength but only moderately
13 impact arm strength. We hypothesized that the mental benefits of exercise might be similarly
14 differentiated. We focused specifically on how different forms of exercise might relate to different
15 aspects of memory and mental health. To test our hypothesis, we collected nearly a century's
16 worth of fitness data (in aggregate). We then asked participants to fill out surveys asking them
17 to self-report on different aspects of their mental health. We also asked participants to engage in
18 a battery of memory tasks that tested their short and long term episodic, semantic, and spatial
19 memory. We found that participants with similar exercise habits and fitness profiles tended to
20 also exhibit similar mental health and task performance profiles.

²¹ **Introduction**

²² Engaging in physical activity (exercise) can improve our physical fitness by increasing muscle
²³ strength (Crane et al., 2013; Knuttgen, 2007; Lindh, 1979; Rogers and Evans, 1993), increasing bone
²⁴ density (Bassey and Ramsdale, 1994; Chilibeck et al., 2012; Layne and Nelson, 1999), increasing
²⁵ cardiovascular performance (Maiorana et al., 2000; Pollock et al., 2000), increasing lung capac-
²⁶ ity (Lazovic-Popovic et al., 2016) (although see Roman et al., 2016), increasing endurance (Wilmore
²⁷ and Knuttgen, 2003), and more. Exercise can also improve mental health (Basso and Suzuki, 2017;
²⁸ Callaghan, 2004; Deslandes et al., 2009; Mikkelsen et al., 2017; Paluska and Schwenk, 2000; Raglin,
²⁹ 1990; Taylor et al., 1985) and cognitive performance (Basso and Suzuki, 2017; Brisswalter et al.,
³⁰ 2002; Chang et al., 2012; Ettnier et al., 2006).

³¹ The physical benefits of exercise can be explained by stress-responses of the affected body tis-
³² sues. For example, skeletal muscles that are taxed during exercise exhibit stress responses (Morton
³³ et al., 2009) that can in turn affect their growth or atrophy (Schiaffino et al., 2013). By comparison,
³⁴ the benefits of exercise on mental health are less direct. For example, one hypothesis is that ex-
³⁵ ercise leads to specific physiological changes, such as increased aminergic synaptic transmission
³⁶ and endorphin release, which in turn act on neurotransmitters in the brain (Paluska and Schwenk,
³⁷ 2000).

³⁸ Speculatively, if different exercise regimens lead to different neurophysiological responses, one
³⁹ might be able to map out a spectrum of signalling and transduction pathways that are impacted
⁴⁰ by a given type, duration, and intensity of exercise in each brain region. For example, prior work
⁴¹ has shown that exercise increases acetylcholine levels, starting in the vicinity of the exercised
⁴² muscles (Shoemaker et al., 1997). Acetylcholine is thought to play an important role in memory
⁴³ formation (Palacios-Filardo et al., 2021, e.g., by modulating specific synaptic inputs from entorhinal
⁴⁴ cortex to the hippocampus, albeit in rodents). Given the central role of these medial temporal
⁴⁵ lobe structures play in memory, changes in acetylcholine might lead to specific changes in memory
⁴⁶ formation and retrieval.

⁴⁷ In the present study, we hypothesize that (a) different exercise regimens will have different,

48 quantifiable impacts on cognitive performance and mental health, and that (b) these impacts will
49 be consistant across individuals. To this end, we collected a year of fitness tracking data from
50 each of 113 participants. We then asked each participant to fill out a brief survey in which they
51 self-evaluated several aspects of their mental health. Finally, we ran each participant through a
52 battery of memory tasks, which we used to evaluate their memory performance along several
53 dimensions. We examined the data for potential associations between memory, mental health, and
54 exercise.

55 **Methods**

56 We ran an online experiment using the Amazon Mechanical Turk platform. We collected data
57 about each participant’s fitness and exercise habits, a variety of self-reported measures concerning
58 their mental health, and about their performance on a battery of memory tasks. We mined the
59 dataset for potential associations between memory, mental health, and exercise.

60 **Experiment**

61 **Participants**

62 We recruited experimental participants by posting our experiment as a Human Intelligence Task
63 (HIT) on the Amazon Mechanical Turk platform. We limited participation to Mechanical Turk
64 Workers who had been assigned a “Masters” designation on the platform, given to workers who
65 score highly across several metrics on a large number of HITs, according to a proprietary algorithm
66 managed by Amazon. We further limited our participant pool to participants who self-reported that
67 they were fluent in English and regularly used a Fitbit fitness tracker device. A total of 160 workers
68 accepted our HIT in order to participate in our experiment. Of these, we excluded all participants
69 who failed to log into their Fitbit account (giving us access to their anonymized fitness tracking
70 data), encountered technical issues (e.g., by accessing the HIT using an incompatible browser,
71 device, or operating system), or who ended their participation prematurely, before completing the

72 full study. In all, 113 participants remained that contributed usable data to the study.

73 For their participation, workers received a base payment of \$5 per hour (computed in 15
74 minute increments, rounded up to the nearest 15 minutes), plus an additional performance-based
75 bonus of up to \$5. Our recruitment procedure and study protocol were approved by Dartmouth's
76 Committee for the Protection of Human Subjects.

77 **Gender, age, and race.** Of the 113 participants who contributed usable data, 77 reported their
78 gender as female, 35 as male, and 1 chose not to report their gender. Participants ranged in age
79 from 19–68 years old (25th percentile: 28.25 years; 50th percentile: 32 years; 75th percentile: 38
80 years). Participants reported their race as White (90 participants), Black or African American (11
81 participants), Asian (7 participants), Other (4 participants), and American Indian or Alaska Native
82 (3 participants). One participant opted not to report their race.

83 **Languages.** All participants reported that they were fluent in either 1 and 2 languages (25th
84 percentile: 1; 50th percentile: 1; 75th percentile: 1), and that they were "familiar" with between 1
85 and 11 languages (25th percentile: 1; 50th percentile: 2; 75th percentile: 3).

86 **Reported medical conditions and medications.** Participants reported having and/or taking med-
87 ications pertaining to the following medical conditions: anxiety or depression (4 participants),
88 recent head injury (2 participants), high blood pressure (1 participant), bipolar (1 participant),
89 hypothyroidism (1 participant), and other unspecified medications (1 participant). Participants
90 reported their current and typical stress levels on a Likert scale as very relaxed (-2), a little relaxed
91 (-1), neutral (0), a little stressed (1), or very stressed (2). The "current" stress level reflected par-
92 ticipants' stress at the time they participated in the experiment. Their responses ranged from -2
93 to 2 (current stress: 25th percentile: -2; 50th percentile: -1; 75th percentile: 1; typical stress: 25th
94 percentile: 0; 50th percentile: 1; 75th percentile: 1). Participants also reported their current level of
95 alertness on a Likert scale as very sluggish (-2), a little sluggish (-1), neutral (0), a little alert (1),
96 or very alert (2). Their responses ranged from -2 to 2 (25th percentile: 0; 50th percentile: 1; 75th
97 percentile: 2). Nearly all (111 out of 113) participants reported that they had normal color vision,

98 and 15 participants reported uncorrected visual impairments (including dyslexia and uncorrected
99 near- or far-sightedness).

100 **Residence and level of education.** Participants reported their residence as being located in the
101 suburbs (36 participants), a large city (30 participants), a small city (23 participants), rural (14 partic-
102 ipants), or a small town (10 participants). Participants reported their level of education as follows:
103 College graduate (42 participants), Master's degree (23 participants), Some college (21 partici-
104 pants), High school graduate (9 participants), Associate's degree (8 participants), Other graduate
105 or professional school (5 participants), Some graduate training (3 participants), or Doctorate (2
106 participants).

107 **Reported water and coffee intake.** Participants reported the number of cups of water and coffee
108 they had consumed prior to accepting the HIT. Water consumption ranged from 0–6 cups (25th
109 percentile: 1; 50th percentile: 3; 75th percentile: 4). Coffee consumption ranged from 0–4 cups (25th
110 percentile: 0; 50th percentile: 1; 75th percentile: 2).

111 **Tasks**

112 Upon accepting the HIT posted on Mechanical Turk, the worker was directed to read and fill out
113 a screening and consent form, and to share access to their anonymized Fitbit data via their Fitbit
114 account. After consenting to participate and successfully sharing their Fitbit data, participants
115 filled out a survey and then engaged in a series of memory tasks (Fig. 1). All stimuli and code for
116 running the full Mechanical Turk experiment may be found [here](#).

117 **Survey questions.** We collected the following demographic information from each participant:
118 their birth year, gender, highest (academic) degree achieved, race, language fluency, and language
119 familiarity. We also collected information about participants' health and wellness, including about
120 their vision, alertness, stress, sleep, coffee and water consumption, location of their residence,
121 activity typically required for their job, and exercise habits.

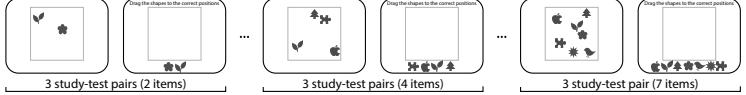
	Main task and immediate memory test				Delayed memory test
a.	1 Free recall	Study words 	Memory test 		5 
b.	2 Naturalistic recall	Watch a short video (The Temple of Knowledge) 	Memory tests  Free response Multiple choice	6 	Free response
c.	3 Foreign language flashcards	Study flashcards 	Memory test 	7 	
d.	4 Spatial learning	Memorize the positions of increasing numbers of shapes 			N/A

Figure 1: Battery of memory tasks. **a. Free recall.** Participants study 16 words (presented one at a time), followed by an immediate memory test where they type each word they remember from the just-studied list. In the delayed memory test, participants type any words they remember studying, from any list. **b. Naturalistic recall.** Participants watch a brief video, followed by two immediate memory tests. The first test asks participants to write out what happened in the video. The second test has participants answer a series of multiple choice questions about the conceptual content of the video. In the delayed memory test, participants (again) write out what happened in the video. **c. Foreign language flashcards.** Participants study a sequence of 10 English-Gaelic word pairs, each presented with an illustration of the given word. During an immediate memory test, participants perform a multiple choice test where they select the Gaelic word that corresponds to the given photograph. During the delayed memory test, participants perform a second multiple choice test, where they select the Gaelic word that corresponds to each of a new set of photographs. **d. Spatial learning.** In each trial, participants study a set of randomly positioned shapes. Next, the shapes' positions are altered, and participants are asked to drag the shapes back to their previous positions. **All panels.** The gray numbers denote the order in which participants experienced each task or test.

¹²² **Free recall (Fig. 1a).** Participants studied a sequence of four word lists, each comprising 16 words.
¹²³ After studying each list, participants received an immediate memory test, whereby they were asked
¹²⁴ to type (one word at a time) any words they remembered from the just-studied list, in any order.

¹²⁵ Words were presented for 2 s each, in black text on a white background, followed by a 2 s blank
¹²⁶ (white) screen. After the final 2 s pause, participants were given 90 s to type in as many words
¹²⁷ as they could remember, in any order. The memory test was constructed such that the participant
¹²⁸ could only see the text of the current word they were typing; when they pressed any non-letter
¹²⁹ key, the current word was submitted and the text box they were typing in was cleared. This was
¹³⁰ intended to prevent participants from retroactively editing their previous responses.

¹³¹ The word lists participants studied were drawn from the categorized lists reported in Ziman
¹³² et al. (2018). Each participant was assigned four unique randomly chosen lists (in a randomized
¹³³ order), selected from a full set of 16 lists. Each chosen list was then randomly shuffled before
¹³⁴ presenting the words to the participants.

¹³⁵ Participants also performed a final delayed memory test where they were given 180 s to type
¹³⁶ out any words they remembered from *any* of the 4 lists they had studied.

¹³⁷ Recalled words within an edit distance of 2 (i.e., a Levenshtein Distance less than or equal to
¹³⁸ 2) of any word in the wordpool were “autocorrected” to their nearest match. We also manually
¹³⁹ corrected clear typos or misspellings by hand (e.g., we corrected “hippopumas” to “hippopota-
¹⁴⁰ mus”, “zucinni” to “zucchini”, and so on). Finally, we lemmatized each submitted word to match
¹⁴¹ the plurality of the matching wordpool word (e.g., “bongo” was corrected to “bongos”, and so
¹⁴² on). After applying these corrections, any submitted words that matched words presented on the
¹⁴³ just-studied list were tagged as “correct” recalls, and any non-matching words were discarded
¹⁴⁴ as “errors.” Because participants were not allowed to edit the text they entered, we chose not to
¹⁴⁵ analyze these putative “errors,” since we could not distinguish typos from true misrememberings.

¹⁴⁶ **Naturalistic recall (Fig. 1b).** Participants watched a 2.5 minute video clip entitled “The Temple
¹⁴⁷ of Knowledge.” The video comprises an animated story told to StoryCorps by Ronald Clark, who
¹⁴⁸ was interviewed by his daughter, Jamilah Clark. The narrator (Ronald) discusses growing up

149 living in an apartment over Washington Heights branch of the New York Public Library, where his
150 father worked as a custodian during the 1940s.

151 After watching the video clip, participants were asked to type out anything they remembered
152 about what happened in the video. They typed their responses into a text box, one sentence at a
153 time. When the participant pressed the return key or typed any final punctuation mark (".", "!", or
154 "?") the text currently entered into the box was "submitted" and added to their transcript, and the
155 text box was cleared to prevent further editing of any already-submitted text. This was intended to
156 prevent participants from retroactively editing their previous responses. Participants were given
157 up to 10 minutes to enter their responses. After 4 minutes participants were given the option of
158 ending the response period early, e.g., if they felt they had finished entering all of the information
159 they remembered. Each participant's transcript was constructed from their submitted responses by
160 combining the sentences into a single document and removing extraneous whitespace characters.

161 Following this 4–10 minute free response period, participants were given a series of 10 multiple
162 choice questions about the conceptual content of the story. All participants received the same
163 questions, in the same order.

164 Participants also performed a final delayed memory test, where they carried out the free
165 response recall task a second time, near the end of the testing session. This resulted in a second
166 transcript, for each participant.

167 **Foreign language flashcards (Fig. 1c).** Participants studied a series of 10 English-Gaelic word
168 pairs in a randomized order. We selected the Gaelic language both for its relatively small number of
169 native speakers and for its dissimilarity to other commonly spoken languages amongst Mechanical
170 Turk Workers. We verified (via self report) that all of our participants were fluent in English and
171 that they were neither fluent nor familiar with Gaelic.

172 Each word's "flashcard" comprised a cartoon depicting the given word, the English word or
173 phrase in lowercase text (e.g., "the boy"), and the Gaelic word or phrase in uppercase text (e.g.,
174 "BUACHAILL"). Each flashcard was displayed for 4 s, followed by a 3 s interval (during which
175 the screen was cleared) prior to the next flashcard presentation.

176 After studying all 10 flashcards, participants were given a multiple choice memory test where
177 they were shown a series of novel photographs, each depicting one of the 10 words they had
178 learned. They were asked to select which (of 4 unique options) Gaelic word went with the given
179 picture. The 3 incorrect options were selected at random (with replacement across trials), and the
180 order in which the choices appeared to the participant were also randomized. Each of the 10 words
181 they had learned were tested exactly once.

182 Participants also performed a final delayed memory test, where they were given a second set of
183 10 questions (again, one per word they had studied). For this second set of questions participants
184 were prompted with a new set of novel photographs, and new randomly chosen incorrect choices
185 for each question. Each of the 10 original words they had learned were (again) tested exactly once
186 during this final memory test.

187 **Spatial learning (Fig. 1d).** Participants performed a series of study-test trials where they memo-
188 rized the onscreen spatial locations of two or more shapes. During the study phase of each trial,
189 a set of shapes appeared on the screen for 10 s, followed by 2 s of blank (white) screen. During the
190 test phase of each trial, the same shapes appeared onscreen again, but this time they were vertically
191 aligned and sorted horizontally in a random order. Participants were instructed to drag (using the
192 mouse) each shape to its studied position, and then to click a button to indicate that the placements
193 were complete.

194 In different study-test trials, participants learned the locations of different numbers of shapes
195 (always drawn from the same pool of 7 unique shapes, where each shape appeared at most one
196 time per trial). They first performed three trials where they learned the locations of 2 shapes; next
197 three trials where they learned the locations of 3 shapes; and so on until their last three trials, where
198 (during each trial) they learned the locations of 7 shapes. All told, each participant performed 18
199 study-test trials of this spatial learning task (3 trials for each of 2, 3, 4, 5, 6, and 7 shapes).

200 **Fitness tracking using Fitbit devices**

201 To gain access to our study, participants provided us with access to all data associated with their
202 Fitbit account from the year (365 calendar days) up to and including the day they accepted the HIT.
203 We filtered out all identifiable information (e.g., participant names, GPS coordinates, etc.) prior to
204 importing their data.

205 **Collecting and processing Fitbit data**

206 The fitness tracking data associated with participants' Fitbit accounts varied in scope and duration
207 according to which device the participant owned (Fig. S1), how often the participant wore (and/or
208 synced) their tracking device, and how long they had owned their device. For example, while all
209 participants' devices supported basic activity metrics such as daily step counts, only a subset of
210 the devices with heart rate monitoring capabilities provided information about workout intensity,
211 resting heart rate, and other related measures.

212 Across all devices, we collected the following information: heart rate data, sleep tracking data,
213 logged bodyweight measurements, logged nutrition measurements, Fitbit account and device
214 settings, and activity metrics.

215 **Heart rate.** If available, we extracted all heart rate data collected by participants' Fitbit device(s)
216 and associated with their Fitbit profile. Depending on the specific device model(s) and settings, this
217 included second-by-second, minute-by-minute, daily summary, weekly summary, and/or monthly
218 summary heart rate information. These summaries include information about participants' aver-
219 age heart rates, and the amount of time they were estimated to have spent in different "heart rate
220 zones" (rest, out-of-range, fat burn, cardio, or peak, as defined by their Fitbit profile), as well as an
221 estimate of the number of estimated calories burned while in each heart rate zone.

222 **Sleep.** If available, we extracted all sleep data collected by participants' Fitbit device(s). Depend-
223 ing on the specific device model(s) and settings, this included nightly estimates of the duration
224 and quality of sleep, as well as the amount of time spent in each sleep stage (awake, REM, light, or

225 deep).

226 **Weight.** If available, we extracted any weight-related information affiliated with participants'
227 Fitbit accounts within 1 year prior to enrolling in our study. Depending on their specific device
228 model(s) and settings, this included their weight, body mass index, and/or body fat percentage.

229 **Nutrition.** If available, we extracted any nutrition-related information affiliated with participants'
230 Fitbit accounts within 1 year prior to enrolling in our study. Depending on their specific account
231 settings and usage behaviors, this included a log of the specific foods they had eaten (and logged)
232 over the past year, and the amount of water consumed each day.

233 **Account and device settings.** We extracted any settings associated with participants' Fitbit ac-
234 counts to determine (a) which device(s) and model(s) are associated with their Fitbit account, (b)
235 time(s) when their device(s) were last synced, and (c) battery level(s).

236 **Activity metrics.** If available, we extracted any activity-related information affiliated with par-
237 ticipants' Fitbit accounts within 1 year prior to enrolling in our study. Depending on their specific
238 device model(s) and settings, this included: daily step counts; daily amount of time spent in each
239 activity level (sedentary, lightly active, fairly active, or very active, as defined by their account
240 settings and preferences); daily number of floors climbed; daily elevation change; and daily total
241 distance traveled.

242 **Comparing recent versus baseline measurements.**

243 We were interested in separating out potential associations between *absolute* fitness metrics and
244 *relative* metrics. To this end, in addition to assessing potential raw (absolute) fitness metrics, we
245 also defined a simple measure of recent changes in those metrics, relative to a baseline:

$$\Delta_{R,B}m = \frac{B \sum_{i=1}^R m(i)}{R \sum_{i=R+1}^{R+B} m(i)},$$

246 where $m(i)$ is the value of metric m from $i - 1$ days prior to testing (e.g., $m(1)$ represents the value
247 of m on the day the participant accepted the HIT, and $m(10)$ represents the value of m 9 days prior
248 to accepting the HIT. Unless otherwise noted, we set $R = 7$ and $B = 30$. In other words, to estimate
249 recent changes in any metric m , we divided the average value of m taken over the prior week by
250 the average value of m taken over the 30 days before that.

251 **Exploratory correlation analyses**

252 We used a bootstrap procedure to identify reliable correlations between different memory-related,
253 fitness-related, and demographic-related variables. For each of $N = 1000$ iterations, we selected
254 (with replacement) a sample of 113 participants to include. This yielded, for each iteration, a
255 sampled “data matrix” with one row per sampled participant and one column for each measured
256 variable. When participants were sampled multiple times in a given iteration, as was often the case,
257 this matrix contained duplicate rows. We used a round-robin imputation procedure to estimate the
258 values of any missing features (Buck, 1960). Next, we computed the Pearson’s correlation between
259 each pair of columns. This yielded, for each pair of columns, a distribution of N bootstrapped
260 correlation coefficients. If fewer than 95% of the coefficients for a given pair of columns had the
261 same sign, we excluded the pair from further analysis and considered the expected correlation
262 between those columns to be undefined. If $\geq 95\%$ of the coefficients for a given pair of columns
263 had the same sign, we computed the expected correlation coefficient as:

$$\mathbb{E}_{i,j}[r] = \tanh\left(\frac{1}{N} \sum_{n=1}^N \tanh^{-1}(\text{corr}(m(i)_n, m(j)_n))\right),$$

264 where $m(x)_n$ represents column x of the bootstrapped data matrix for iteration n , \tanh is the
265 hyperbolic tangent, and \tanh^{-1} is the inverse hyperbolic tangent.

266 **Regression-based prediction analyses**

267 Following our exploratory correlation analyses, we used an analogous bootstrap procedure to iden-
268 tify subsets of memory-related, fitness-related, and demographic-related variables that predicted

269 (non-overlapping) subsets of other variables. For example, we tested whether a combination of
270 fitness-related variables could predict a combination of memory-related variables, and so on.

271 We used the same bootstrap procedure described above (used in our exploratory correlation
272 analyses) to generate $N = 1000$ bootstrapped data matrices whose rows reflected sampled partici-
273 pants and whose columns reflected different measured variables.

274 We grouped variables according to whether they were memory-related, fitness-related, or
275 demographic-related. For each bootstrap iteration, we divided the rows of that iterations data
276 matrix into training and test sets. The assignments of rows to these two sets was random, subject
277 to the constraint that any duplicated rows in the data matrix (i.e., reflecting a single participant who
278 had been sampled multiple times) was always assigned to either the training *or* the test set—i.e.,
279 duplicated rows could not appear in both the training and the test sets. The training sets always
280 comprised 75% of the data, and the tests sets comprised the remaining 25% of the data.

281 Next, we fit a series of ridge regression models to the training data. Specifically, for each pairing
282 of memory, fitness, and demographic variables, we fit a single ridge regression model treating the
283 first variable group as the input features and the second variable group as the target features. For
284 example, one regression model used memory variables to predict fitness variables, and another
285 regression model used fitness variables to predict demographic variables, and so on. In total we
286 fit six regression models to each training dataset. We then applied the fitted models to the held-
287 out test dataset and computed the root mean squared deviation (RMSD) between the predicted
288 and observed values in the target features of the test dataset. We also examined the regression
289 weights assigned to each input feature. This yielded, for each regression model (across N bootstrap
290 iterations) a distribution of RMSD values and a distribution of weights for each input variable.

291 We constructed a “null” distribution by using the same procedure as above, but where the
292 columns in the test datasets were randomly permuted with each iteration (thereby breaking any
293 meaningful predictive information between the training and test data). We assessed the statistical
294 significance (p -values) of the observed RMSD values by computing the proportions of null RMSD
295 values that were less than the observed value. We also assessed the significance of the observed
296 regression weights using t -tests to compare the means of the observed versus null distributions of

297 weights.

298 **Reverse correlation analyses**

299 We sought to characterize potential associations between the history of participants' fitness-related
300 activities leading up to the time they participated in a memory task and their performance on
301 the given task. For each fitness-related variable, we constructed a timeseries matrix whose rows
302 corresponded to timepoints (sampled once per day) leading up to the day the participant accepted
303 the HIT for our study, and whose columns corresponded to different participants. These matrices
304 often contained missing entries, since different participants' Fitbit devices tracked fitness-related
305 activities differently. For example, participants whose Fitbit devices lacked heart rate sensors
306 would have missing entries for any heart rate-related variables. Or, if a given participant neglected
307 to wear their fitness tracker on a particular day, the column corresponding to that participant
308 would have missing entries for that day.

309 In addition to this set of matrices storing timeseries data for each fitness-related variable, we also
310 constructed a memory performance matrix, M , whose rows corresponded to different memory-
311 related variables, and whose columns corresponded to different participants. For example, one
312 row of the memory performance matrix reflected the average proportion of words (across lists)
313 that each participant remembered during the immediate free recall test, and so on.

314 Given a fitness timeseries matrix, F , we computed the weighted average and weighted standard
315 error of the mean of each row of F , where the weights were given by a particular memory-related
316 variable (row of M). For example, if F contained participants' daily step counts, we could use
317 any row of M to compute a weighted average across any participants who contributed step count
318 data on each day. Choosing a row of M that corresponded to participants' performance on the
319 naturalistic recall task would mean that participants who performed better on the naturalistic recall
320 task would contribute more to the weighted average timeseries of daily step counts. Specifically,

321 for each row, t , of F , we computed the weighted average (across the S participants) as:

$$\bar{f}(t) = \sum_{s=1}^S \hat{m}(s)F(t,s),$$

322 where \hat{m} denotes the normalized min-max scaling of m (the row of M corresponding to the chosen
323 memory-related variable):

$$\hat{m} = \frac{m}{\sum_{s=1}^S \hat{m}(s)},$$

324 where

$$\hat{m} = \frac{m - \min(m)}{\max(m) - \min(m)}$$

325 We computed the weighted standard error of the mean as:

$$\text{SEM}_m(f(t)) = \frac{\left| \sum_{s=1}^S (F(t,s) - \bar{f}(t)) \right|}{\sqrt{S}}.$$

326 When a given row of F was missing data from one or more participants, those participants were
327 excluded from the weighted average for the corresponding timepoint and the weights (across all
328 remaining participants) were re-normalized to sum to 1. The above procedure yielded, for each
329 memory variable, a timeseries of average (and standard error of the mean) fitness tracking values
330 leading up to the day of the experiment.

331 Results

332 Before testing our main hypothesis we examined the behavioral data from each of four memory
333 tasks: a random word list learning “free recall” task; a naturalistic recall task whereby participants
334 watched a short video and then recounted the narrative; a foreign language “flashcards” task; and
335 a spatial learning task. Each of the first three tasks (free recall, naturalistic recall, and the flashcards
336 task) included both an immediate (short term) memory test and a delayed (long term) memory test.
337 The spatial learning task included only an immediate test. Participants in all four tasks exhibited

338 general trends and tendencies that have been previously reported in prior work. We were also
339 interested in characterizing the variability in task performance across participants. For example,
340 if all participants exhibited near-identical behaviors or performance on a given task, we would be
341 unable to identify how memory performance on that task varied with mental health or exercise.

342 When participants engaged in free recall of random word lists, they displayed strong primacy
343 and recency effects (Murdock, 1962) on the immediate memory tests (as reflected by improved
344 memory for early and late list items; Fig. 2a, left and right panels). On the delayed memory test,
345 the recency effect was substantially diminished (Fig. 3a, left and right panels), consistent with
346 myriad previous studies (for review see Kahana, 2012). Participants also tended to cluster their
347 recalls according to the words' study positions (Kahana, 1996) on both the immediate (Fig. 2a,
348 middle panel) and delayed (Fig. 3a, middle panel) memory tests.

349 When participants engaged in naturalistic recall by recounting the narrative of a short story
350 video, they reliably and accurately remembered the major narrative events on both the immediate
351 (Fig. 2b) and delayed (Fig. 3b) tests. This is consistent with prior work showing that memory for
352 rich narratives is both detailed and accurate (Chen et al., 2017; Heusser et al., 2021).

353 Performance on the foreign language flashcards task (immediate: Fig. 2c; delayed: Fig. 3c)
354 varied substantially across participants, and did not show any clear serial position effects. Participants
355 also displayed substantial variation in performance on the spatial learning task (Fig. 2d).
356 In general, participants reported the shape's positions more accurately when there were fewer
357 shapes. However, both the baseline estimation accuracy and the rate of decrease in accuracy as a
358 function of increasing number of memorized locations varied substantially across participants.

359 In addition to observing substantial across-participant variability in memory performance,
360 we also observed substantial variability in participants' fitness and activity metrics (Fig. 4). We
361 examined recent measurements, averaged over the week prior to testing (Fig. 4a), baselined mea-
362 surements (average over the prior week, divided by the average over the preceding 30 day; Fig. 4b),
363 along with more gradually varying measures that tended to remain relatively static over timescales
364 of weeks to months (Fig. 4c). Figure S6 displays across-participant distributions for a broad selec-
365 tion of these measures, and Figures S7, S8, S9, and S10 show different participants' fitness metrics,

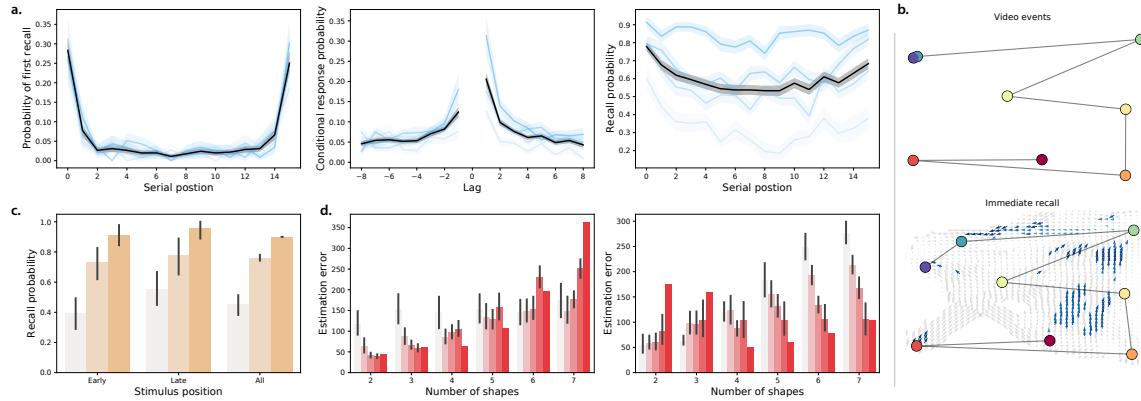


Figure 2: Immediate memory tests. **a. Free recall.** Left: probability of recalling each word first as a function of its presentation position. Middle: probability of transitioning between successively recalling the word presented at position i , followed by word presented at position $i + \text{Lag}$. Right: probability of recalling each word as a function of its presentation position. See Figure S2 for additional details. **b. Naturalistic recall.** Top: 2D embedding of a 2.5 min video clip; each dot reflects a narrative event (red denotes early events and blue denotes later events). Bottom: 2D embedding of the averaged transcripts of participants' recounts (dots: same format as top panel). The arrows denote the average trajectory directions through the corresponding region of text embedding space, for any participants whose recounts passed through that region. Blue arrows denote statistically reliable agreement across participants ($p < 0.05$, corrected). See Figure S3 for additional details. **c. Foreign language flashcards.** Each bar denotes the average proportion of correctly recalled Gaelic-English word pairs from early (first 3), late (last 3), or all (i.e., all 10) study positions. See Figure S4 for additional details. **d. Spatial learning.** Average estimation error in shape locations as a function of the number of shapes. See Figure S5 for additional details. All panels: error bars and error ribbons denote bootstrap-estimated 95% confidence intervals. Shading (saturation) denotes results for different subsets of participants assigned based on their task performance (Figs. S2, S3, S4, and S5 provide information about which performance metrics and values the shading reflects; in general more saturated colors denote participants who performed better on the given task.) In Panel d, participants are grouped in two ways; in the left panel, participants are grouped according to the y -intercepts of regression lines (estimation error as a function of the number of shapes); in the right panel, participants are grouped according to the slopes of the same regression lines.

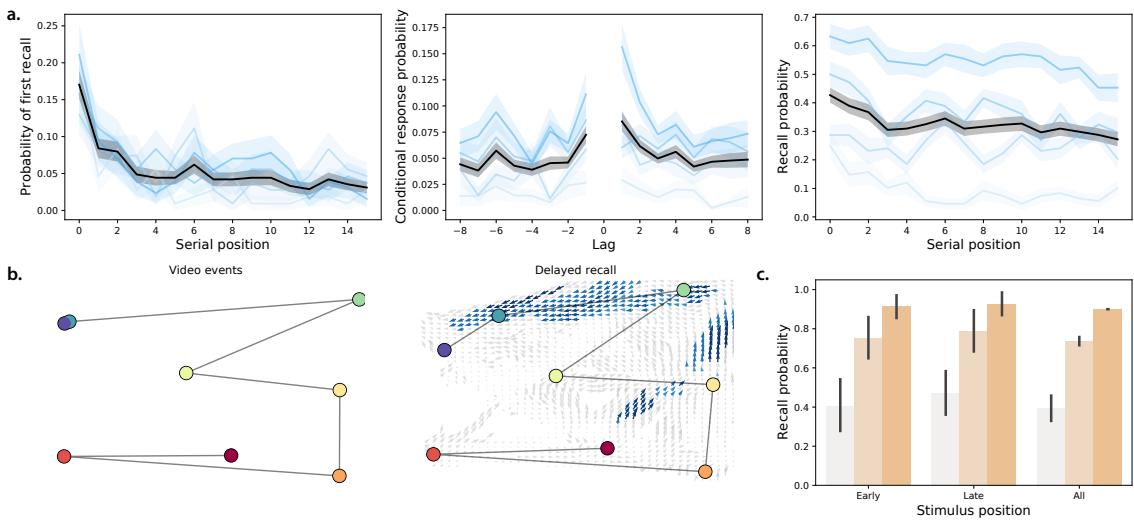


Figure 3: Delayed memory tests. **a. Free recall.** These panels are in the same format as Figure 2a, but they reflect performance on the delayed free recall task. For additional details see Figure S2. **b. Naturalistic recall.** These panels are in the same format as Figure 2b, but the right panel reflects performance on the delayed naturalistic recall task. For additional details see Figure S3. **c. Foreign language flashcards.** This panel is in the same format as Figure 2c, but it reflects performance on the delayed flashcards test. For additional details see Figure S4.

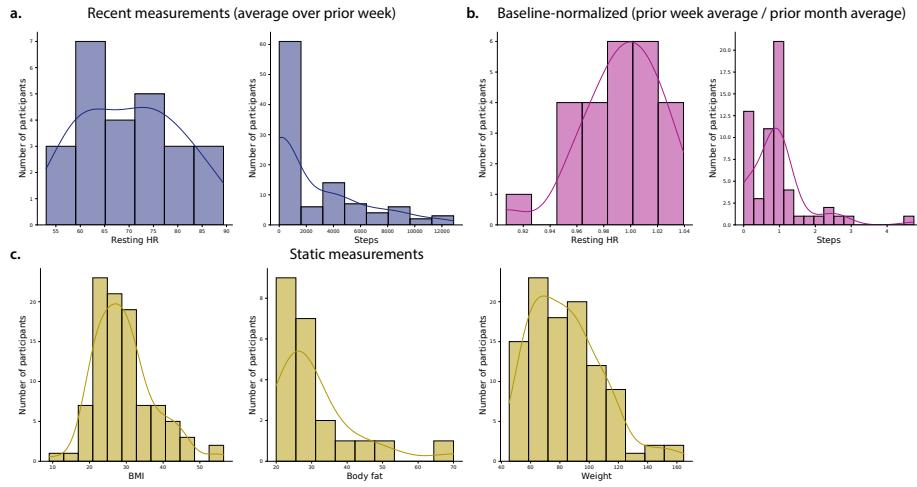


Figure 4: Fitness measures. **a. Recent measures.** Resting heart rate (HR) and daily step counts, averaged over the week prior to testing. **Baseline-normalized measures.** Resting heart rate and daily step counts averaged over the week prior to testing, divided by the average resting heart rate and step counts averaged over the preceding month. **Static measures.** Body mass index (BMI), body fat percentage, and weight (in kg). For more information see Figures S6, S7, S8, S9, and S10.

366 broken down by their performance on different memory tasks.

- 367 • exploratory analysis (correlations). Possibly make some sort of scatter matrix or pairplot-
368 rows/columns: tasks. Diagonal: histogram of performance metric for that task. Above-
369 diagonal entries: compare performance across tasks (by subject). Below-diagonal entries:
370 empty? density/2d histograms?
- 371 – Memory-memory
- 372 – fitness-fitness
- 373 – survey-survey
- 374 – (fitness + survey)-memory
- 375 • predictive analysis (regressions)
 - 376 – Predict memory performance on held-out task from other tasks
 - 377 – Predict memory performance on each task using fitness data

- 378 – Predict memory performance on each task using survey data
- 379 • Reverse correlations: look at recent changes versus baseline trends (color using same scheme
- 380 as behavior figures). Possibly
- 381 – Fitness profile that predicts performance on each task (barplots + timelines)
- 382 – Fitness profile for each survey demographic (barplots + timelines)
- 383 * Select out mental health demographics (based on meds, stress levels)

384 **Discussion**

- 385 • summarize key findings
- 386 • correlation versus causation
- 387 • what can vs. can't we know? we can identify correlations, but not causal direction– e.g. we
- 388 cannot know whether exercise *causes* mental changes versus whether people with particular
- 389 neural profiles might tend to engage in particular exercise behaviors. that being said, we *can*
- 390 separate out baseline tendencies (e.g., how people tend to exercise in general) versus recent
- 391 changes (e.g., how they happened to have exercised prior to the experiment).
- 392 • related work (exercise/memory, exercise/mental health), what this study adds
- 393 • future direction: towards customized physical exercise recommendation engine for optimiz-
- 394 ing mental health and mental fitness

395 **Acknowledgements**

396 We acknowledge useful discussions with David Bucci, Emily Glasser, Andrew Heusser, Abigail

397 Bartolome, Lorie Loeb, Lucy Owen, and Kirsten Ziman. Our work was supported in part by

398 the Dartmouth Young Minds and Brains initiative. The content is solely the responsibility of the

399 authors and does not necessarily represent the official views of our supporting organizations. This

400 paper is dedicated to the memory of David Bucci, who helped to inspire the theoretical foundations
401 of this work. Dave served as a mentor and colleague on the project prior to his passing.

402 **Data and code availability**

403 All analysis code and data used in the present manuscript may be found [here](#).

404 **Author contributions**

405 Concept: J.R.M. Experiment implementation and data collection: G.M.N. Analyses: G.M.N., E.C.,
406 P.C.F., and J.R.M. Writing: J.R.M.

407 **Competing interests**

408 The authors declare no competing interests.

409 **References**

- 410 Bassey, E. J. and Ramsdale, S. J. (1994). Increase in femoral bone density in young women following
411 high-impact exercise. *Osteoporosis International*, 4:72–75.
- 412 Basso, J. C. and Suzuki, W. A. (2017). The effects of acute exercise on mood, cognition, neurophys-
413 iology, and neurochemical pathways: a review. *Brain Plasticity*, 2(2):127–152.
- 414 Brisswalter, J., Collardeau, M., and René, A. (2002). Effects of acute physical exercise characteristics
415 on cognitive performance. *Sports Medicine*, 32:555–566.
- 416 Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use
417 with an electronic computer. *Journal of the Royal Statistical Society*, 22(2):302–306.

- 418 Callaghan, P. (2004). Exercise: a neglected intervention in mental health care? *Psychiatric and*
419 *Mental Health Nursing*, 11(4):476–483.
- 420 Chang, Y. K., Labban, J. D., Gapin, J. I., and Etnier, J. L. (2012). The effects of acute exercise on
421 cognitive performance: a meta-analysis. *Brain Research*, 1453:87–101.
- 422 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared
423 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,
424 20(1):115.
- 425 Chilibek, P. D., Sale, D. G., and Webber, C. E. (2012). Exercise and bone mineral density. *Sports*
426 *Medicine*, 19:103–122.
- 427 Crane, J. D., MacNeil, L. G., and Tarnopolsky, M. A. (2013). Long-term aerobic exercise is associated
428 with greater muscle strength throughout the life span. *The Journals of Gerontology: Series A*,
429 68(6):631–638.
- 430 Deslandes, A., Moraes, H., Ferreira, C., Veiga, H., Silveira, H., Mouta, R., Pompeu, F. A. M. S.,
431 Coutinho, E. S. F., and Laks, J. (2009). Exercise and mental health: many reasons to move.
432 *Neuropsychobiology*, 59:191–198.
- 433 Etnier, J. L., Nowell, P. M., Landers, D. M., and Sibley, B. A. (2006). A meta-regression to examine the
434 relationship between aerobic fitness and cognitive performance. *Brain Research: Brain Research*
435 *Reviews*, 52(1):119–130.
- 436 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behavioral
437 and neural signatures of transforming naturalistic experiences into episodic memories. *Nature*
438 *Human Behavior*, In press.
- 439 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*, 24:103–109.
- 440 Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press, New York, NY.
- 441 Knutgen, H. G. (2007). Strength training and aerobic exercise: comparison and contrast. *Journal of*
442 *Strength and Conditioning Research*, 21(3):973–978.

- 443 Layne, J. E. and Nelson, M. E. (1999). The effects of progressive resistance training on bone density:
444 a review. *Medicine and Science in Sports and Exercise*, 31(1):25–30.
- 445 Lazovic-Popovic, B., Zlatkovic-Svenda, M., Durmic, T., Djelic, M., Saranovic, D., and Zugic, V.
446 (2016). Superior lung capacity in swimmers: some questions, more answers! *Revista Portuguesa
447 de Pneumologia*, 22(3):151–156.
- 448 Lindh, M. (1979). Increase of muscle strength from isometric quadriceps exercises at different knee
449 angles. *Scandinavian Journal of Rehabilitation Medicine*, 11(1):33–36.
- 450 Maiorana, A., O'Driscoll, G., Cheetham, C., Collis, J., Goodman, C., Rankin, S., Taylor, R., and
451 Green, D. (2000). Combined aerobic and resistance exercise training improves functional capacity
452 and strength in CHF. *Journal of Applied Physiology*, 88(1565–1570).
- 453 Mikkelsen, K., Stojanovska, L., Polenakovic, M., Bosevski, M., and Apostolopoulos, V. (2017).
454 Exercise and mental health. *Maturitas*, 106:48–56.
- 455 Morton, J. P., Kayani, A. C., McArdle, A., and Drust, B. (2009). The exercise-induced stress response
456 of skeletal muscle, with specific emphasis on humans. *Sports Medicine*, 39:643–662.
- 457 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology:
458 General*, 64:482–488.
- 459 Palacios-Filardo, J., Udakis, M., Brown, G. A., Tehan, B. G., Congreve, M. S., Nathan, P. J., Brown, A.
460 J. H., and Mellor, J. R. (2021). Acetylcholine prioritises direct synaptic inputs from entorhinal cor-
461 tex to CA1 by differential modulation of feedforward inhibitory circuits. *Nature Communications*,
462 12(5475):doi.org/10.1038/s41467-021-25280-5.
- 463 Paluska, S. A. and Schwenk, T. L. (2000). Physical activity and mental health. *Sports Medicine*,
464 29(3):167–180.
- 465 Pollock, M. L., Franklin, B. A., Balady, G. J., Chaltman, B. L., Fleg, J. L., Fletcher, B., Limacher, M.,
466 na, I. L. P., Stein, R. A., Williams, M., and Bazzarre, T. (2000). Resistance exercise in individuals
467 with and without cardiovascular disease. *Circulation*, 101:828–833.

- 468 Raglin, J. S. (1990). Exercise and mental health. *Sports Medicine*, 9:323–329.
- 469 Rogers, M. A. and Evans, W. J. (1993). Changes in skeletal muscle with aging: effects of exercise
470 training. *Exercise and Sport Sciences Reviews*, 21:65–102.
- 471 Roman, M. A., Rossiter, H. B., and Casaburi, R. (2016). Exercise, ageing and the lung. *European
472 Respiratory Journal*, 48:1471–1486.
- 473 Schiaffino, S., Dyar, K. A., Ciciliot, S., Blaauw, B., and Sandri, M. (2013). Mechanisms regulating
474 skeletal muscle growth and atrophy. *The febs Journal*, 280(17):4294–4314.
- 475 Shoemaker, J. K., Halliwill, J. R., Hughson, R. L., and Joyner, M. J. (1997). Contributions of
476 acetylcholine and nitric oxide to forearm blood flow at exercise onset and recovery. *Vascular
477 Physiology*, 273(5):2388–2395.
- 478 Taylor, C. B., Sallis, J. F., and Needle, R. (1985). The relation of physical activity and exercise to
479 mental health. *Public Health Reports*, 100(2):195–202.
- 480 Wilmore, J. H. and Knuttgen, H. G. (2003). Aerobic exercise and endurance. *The Physician and
481 Sportsmedicine*, 31(5):45–51.
- 482 Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is automatic
483 speech-to-text transcription ready for use in psychological experiments? *Behavior Research
484 Methods*, 50:2597–2605.